

EDUCATION

- **Carnegie Mellon University School of Computer Science** Pittsburgh, PA
B.S. in Artificial Intelligence; GPA: 3.92/4.00, Dean's List. Relevant courses: Expected Dec 2022
 - **CS:** *Database, Distributed Systems, *Search Engines, Parallel Data Structures & Algorithms, Software Design
 - **ML:** Advanced Deep Learning, *Deep Learning Systems, Natural Language Processing, ML with Large Datasets

WORK EXPERIENCE

- **Uber** San Francisco, CA
Software Engineer Intern Jun 2021 - Aug 2021
 - End-to-end owned, designed, developed, and launched the Uber Eats home feed dish recommendation carousel to 90M global users, boosting top-level business metrics. Coded in Java, Go, PySpark, and HiveQL.
 - Trained and indexed DL embeddings in Uber's homegrown search system. Served embeddings for candidates retrieval using a novel approach that elevated recall rate by 4x with the same resource as baseline.
 - Implemented eater history retrieval based on personalized order and click data.
 - Prepared feature pipelines. Trained, tuned, and served a XGBoost model for candidates ranking.
- **ByteDance (TikTok)** Beijing
Software Engineer Intern Jun 2020 - Aug 2020
 - **Live Stream Recommendation with Graph Embedding:**
 - * Built a distributed user-author graph from petabyte log data using MapReduce.
 - * Devised ML graph encoders and end-to-end training architecture with Tensorflow to predict click-through-rate.
 - * Optimized mini-batch forward latency of internal ML trainer by 40%+ in graph embedding training.
 - * Boosted TikTok online user staytime +3.5%, etc. in AB tests and rolled out to 600M Chinese users.
 - **Systems for Engineering Efficiency:**
 - * Developed a model health monitor and alert system from scratch in Django with RESTful APIs. Onboarded 100+ online models across 5 products with 50+ internal users. Reduced response time to <1hr.
 - * Constructed an analysis pipeline on 300+ features that modifies terabyte model checkpoints distributedly based on analysis result. Saved 35k+ core-hour computing resources than hand-tuning.
- **CMU Machine Learning Department** Pittsburgh, PA
Teaching Assistant for 10-605 Machine Learning with Large Datasets Feb 2021 - Jun 2021
 - Designed a major assignment from scratch, with write-ups, tutorial videos, and starter codes. Onboarded 140+ students to ML at scale with Spark and AWS.
 - Wrote exams; led weekly recitations and office hours for 20+ undergraduate and graduate students.

RESEARCH EXPERIENCE

- **TheSys Group, CMU Parallel Data Lab** Pittsburgh, PA
Research Assistant Nov 2020 - Present
 - Researched fault tolerance in distributed recommender systems training with Prof. Rashmi K. Vinayak.
 - Coded and experimented with different fault tolerance strategies, e.g. replication, checkpointing, and erasure coding, in the open-source training system XDL to understand efficiency tradeoffs.
 - Proposed a novel multi-level approach that utilizes a hybrid fault tolerance strategy to minimize time and memory overhead. Worked on its implementation, benchmarking, and paper drafting.

PROJECTS

- **TinyNet:** (WIP) A PyTorch-like automatic differentiation library with CPU and GPU acceleration. (C++, Python)
- **AlpacaHub:** (WIP) An env, data, and model versioning framework for machine learning workflows. (JS, Python)
- **QASys:** A question generation and answering system on text with rule-based and neural backend. (NLTK, PyTorch)
- **BitcoinMiner:** A failure-recoverable distributed Bitcoin miner with the homegrown Live Sequence Protocol. (Go)
- **Finger:** A tiny-screen-optimized input keyboard with trie and ngram empowered autocompletion. (Java)
- **Pop!:** A crowd-sourcing notification app that allows users to send signals to groups in real-time. (React, Django)

SKILLS

- **Languages:** Java, Go, C/C++, Python, SQL, Standard ML
- **DevOps:** Spark/MapReduce, Tensorflow/PyTorch, Hive/Presto, Docker, HDFS, Kafka, Protobuf, ElasticSearch